OBSCROWD: JOINT FACE-LOCALIZED IDENTITY PROTECTION FOR MULTI-PORTRAIT IMAGES UNDER GENERATIVE EDITING

Anonymous authors
Paper under double-blind review

ABSTRACT

Generative image editors can modify group photographs while preserving biometric cues that allow face-recognition systems to identify the depicted individuals. We present Obscrowd, a method that jointly optimizes a single perturbation localized to all detected faces through soft masks. A diffusion teacher–student pipeline with inference-time refinement optimizes identity disruption across simulated edits using an ensemble of face recognizers. We evaluate Obscrowd on group photographs and single portraits, comparing protected faces with matched unprotected controls to distinguish protection from identity loss caused by editing itself. Evaluation examines both crop–edit orderings, transfer to a held-out recognizer, and the effects of perturbation budget and perceptual cost. Across the evaluated editing and processing conditions, Obscrowd reduces biometric similarity significantly relative to matched unprotected controls. Under crop-then-edit evaluation with an instruction-based editor, mean cosine similarity on group photographs falls from 0.662 to 0.287. Joint optimization also achieves protection comparable to independent per-face optimization using approximately one-quarter of the compute. Alongside these gains, protection is limited to detected faces and weakens as wider crops reduce the protected face's share of the recognizer input. These findings support joint optimization as a computationally efficient approach to protecting multiple faces under generative editing, while identifying cropping and adaptive attacks as priorities for further investigation.

1 INTRODUCTION

Diffusion editors can realistically alter a portrait while preserving biometric cues that let face-recognition (FR) systems link the edited image to its subject. The image may look new yet still betray *who* is depicted.

Two different contracts. A growing family of proactive defenses perturbs an image to make downstream editing fail (Salman et al., 2023; Chen et al., 2024; Wang et al., 2025a; Song et al., 2025; Xu et al., 2025; Ahn et al., 2025; Wang et al., 2025b). That **anti-editing** objective differs from a **privacy** objective: make the edited identity unmatchable even when the edit succeeds. The latter matters for photographs meant to be posted, re-encoded, cropped or edited. If a protected edit remains close to the corresponding unprotected edit, the subject may still be identifiable. We measure this conditional risk in Section 5.5; perceptual closeness alone does not prove that the requested edit succeeded.

Group photographs. Most protections target single-subject portraits or perturb the whole image, while shared photographs often contain several faces. Faces differ in scale, attackers can crop an individual, and a whole-image perturbation spends about 91% of its budget outside the face regions. Protection must hold per face.

Our approach. **Obscrowd** uses soft masks to place one bounded perturbation across all detected faces (Figure 1). A frozen diffusion teacher produces training targets offline; an 8.26M-parameter student predicts the perturbation, which projected-gradient refinement (Madry et al., 2018) adjusts at

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

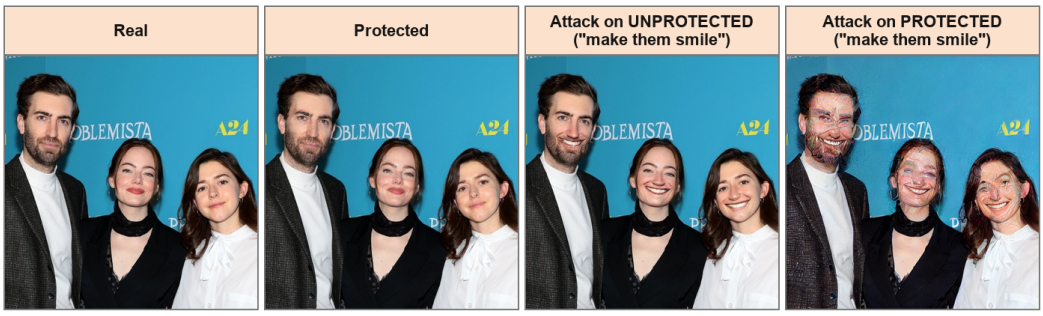

Figure 1: **Joint protection of three faces.** After InstructPix2Pix ("make them smile"), the unprotected edit remains matchable while the protected edit disrupts all three identities. The perturbation is visible at the deployed budget (Section 5.4).

inference time. An FR ensemble measures identity after simulated edits. To evaluate protection, we compare each protected face with the same edit of its unprotected version. We test both sequences: crop the face before editing, or edit the full image before cropping. Section 4 defines the thresholds and perturbation costs used for comparison.

Contributions. We measure residual identity when protected and clean edits remain close; protect all detected faces jointly at $4.2\times$ less compute than independent per-face optimization with comparable strength ($-0.0009\,[-0.005, +0.004]$); compare with PhotoGuard, EditShield and FaceLock under editing, plus IDProtector under identity-conditioned generation, using per-face PGD as a control; and quantify crop dilution ($+0.665$ to $+0.108$).

2 Related Work

Anti-editing immunization. PhotoGuard (Salman et al., 2023) and EditShield (Chen et al., 2024) perturb an image so a downstream editor produces a visibly broken output, and Anti-DreamBooth (Le et al., 2023) pursues the same contract against subject-driven finetuning. We evaluate them on the axis they optimize as well as on identity, and quantify what their contract costs when it is not met (Section 5.5).

Identity-disrupting protection. FaceLock (Wang et al., 2025a) optimizes a whole-image perturbation that destroys biometric features after editing and argues, as we do, for FR-based rather than pixel-based evaluation. IDProtector (Song et al., 2025) trains a feed-forward encoder attacking the ArcFace embedding consumed by identity-conditioned generators such as InstantID (Wang et al., 2024) and IP-Adapter (Ye et al., 2023). GoodAC (Xu et al., 2025), FastProtect (Ahn et al., 2025) and Adv-CPG (Wang et al., 2025b) respectively couple global with local feature disruption, amortize per-image optimization offline, and embed the objective inside the generation pipeline. These share our objective but are developed on single-subject portraits and perturb the whole image, and our positioning is correspondingly narrow: FaceLock's own text names multi-individual images as a remaining challenge, and IDProtector's single-face restriction is structural, its preparation stage sorting detected faces by area and conditioning on one. Sections 5.2 and 5.3 test both on the authors' released code, each method inside the threat model it was built for.

Purification, transfer and efficiency. Purifiers threaten protection by denoising the protective signal away, sharpened with attention-informed schedules (Pei et al., 2025) or distilled to one step (Lei et al., 2025) and motivating anti-purification defenses (Salar et al., 2025; Gan et al., 2025); transferable attacks on FR show a defense tuned to one recognizer can leak under another (Zhou et al., 2025; Li et al., 2025). We claim no defense against an adaptive purifier, and we separate purification the refinement stage modeled from purification it did not, since conflating the two makes robustness claims unreadable. Work amortizing protection with learned encoders (Song et al., 2025; Ahn et al., 2025) or data-free universal perturbations (Lee et al., 2025) shares our student's motivation, but we retain refinement because the generator alone leaves substantial residual identity.

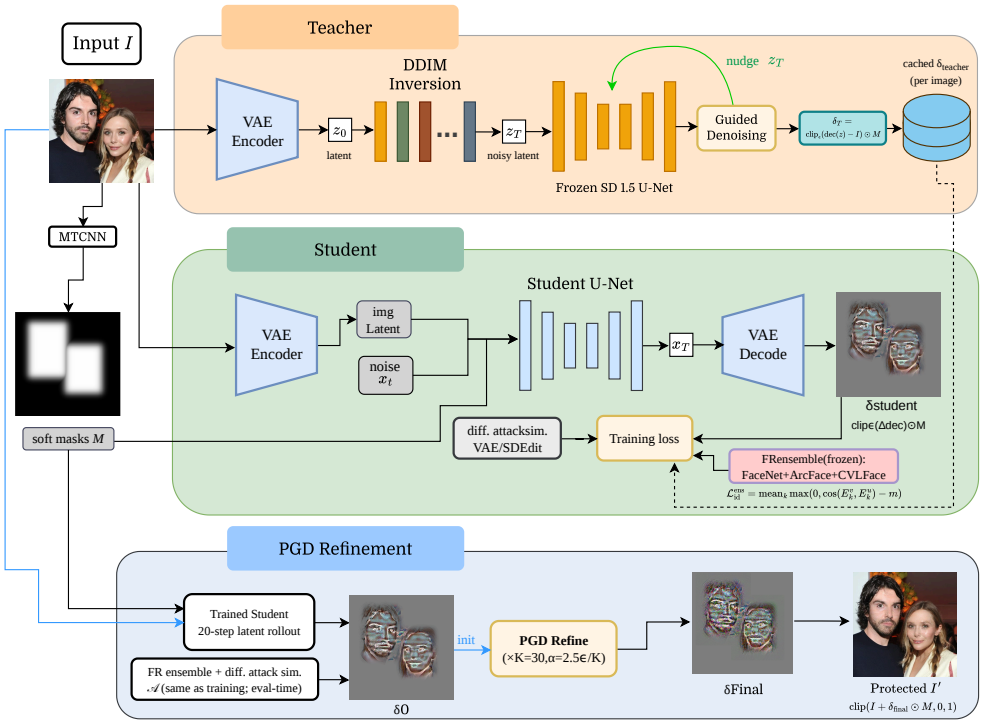

Figure 2: **Obscrowd pipeline.** *Teacher (offline):* a frozen Stable Diffusion 1.5 caches one mask-localized perturbation per image. *Student:* a latent conditional U-Net (8.26M parameters) distills it with an L_1 term plus an identity-disruption term from a frozen FR ensemble evaluated *through* a differentiable attack simulator.

Evaluation methodology. The confounds we control are not specific to our method. Protection strength is routinely reported as a post-attack similarity rather than a difference against a matched control, at a flat threshold rather than a per-recognizer calibrated one, with intervals over faces rather than images, and at a fixed L_∞ cap rather than a matched realized density. Each inflates a reported margin, so Section 4 treats them as part of the contribution. Adjacent threads on diffusion backdoors, attacks on detectors and watermarks, robust detection and inversion bounds (Han et al., 2025; Yuan et al., 2025; Müller et al., 2025; Wang et al., 2025c; Kashiani et al., 2025; Huang et al., 2025; Guo et al., 2025; Xia et al., 2025) address provenance, not multi-face unlinkability under editing.

3 Method

3.1 Problem Formulation and Threat Models

Let $I \in [0,1]^{H\times W\times 3}$ contain N detected faces with soft masks $\{M_i\}_{i=1}^N$, obtained by Gaussian-blurring the detector's boxes so the perturbation has no visible seam, and write $M = \max_i M_i$. We seek one perturbation tensor δ giving

$$I' \;=\; \text{clip}_{[0,1]}\big(I + \delta \odot M\big), \qquad \|\delta\|_\infty \le \epsilon. \tag{1}$$

The same δ serves every face, one optimization for N subjects, and because it is multiplied by M the budget is spent only on pixels a recognizer reads. The student consumes 8 mask channels, so faces are sorted by area and processed in tiles of 8 merged by a per-pixel maximum, leaving N unbounded; truncating instead to the first eight masks would leave 20 of our 1,373 multi-portrait faces unprotected while still counting them in every average (Appendix A.2).

We evaluate under two threat models, because the closest prior works sit in different ones and comparing across them is a category error. In **TM-1, generative editing**, the defender publishes I' once, before knowing the attacker's choices; the attacker applies an operation $\mathcal{A}$, optionally crops a face,

and queries an off-the-shelf recognizer E, so protection succeeds when $\mathrm{sim}\big(E(I_i), E(\mathcal{A}(I')_i)\big)$ falls below E's threshold at a stated false-accept rate. In **TM-2, identity-conditioned generation**, the photograph is instead an identity *reference* for a pipeline conditioning on an ArcFace embedding, so protection succeeds when that embedding diverges. Adaptive purifiers are out of scope in both.

3.2 Teacher: Diffusion-Guided Perturbation

The teacher is a frozen Stable Diffusion 1.5 (Rombach et al., 2022). It encodes I, runs a short DDIM trajectory (Song et al., 2021), and at each step decodes the predicted clean latent z to evaluate

$$\mathcal{L}_{\mathrm{T}} = -\lambda_{\mathrm{id}}\mathcal{L}_{\mathrm{id}} - \lambda_{\mathrm{loc}}\mathcal{L}_{\mathrm{loc}} + \lambda_{\mathrm{perc}}\mathcal{L}_{\mathrm{perc}}, \tag{2}$$

with $\mathcal{L}_{\mathrm{id}}$ the ensemble identity score of Equation (3) on the decoded image, $\mathcal{L}_{\mathrm{loc}}$ the same term on the masked face crops, and $\mathcal{L}_{\mathrm{perc}}$ an LPIPS (Zhang et al., 2018) term against I. The identity scores are non-positive and are maximized toward zero by minimizing Equation (2); the perceptual term is minimized. The decoded difference is projected onto the ϵ-ball and cached, $\delta_{\mathrm{T}} = \Pi_\epsilon(\mathrm{dec}(z) - I) \odot M$. The offline guidance generates per-image training targets without updating the diffusion weights; inference uses only the student and refinement stage.

3.3 Student: Distillation Through a Differentiable Attack

The student G_θ is a conditional U-Net in the SD latent space taking 16 input channels (noisy latent, image latent, eight masks) at base width 64, with 8,257,060 trainable parameters, emitting a pixel-space perturbation bounded by $\epsilon\tanh(\cdot)$ from a DDIM-style rollout (2 steps in training, 20 at inference). With $\{E_1, \ldots, E_R\}$ the frozen FR ensemble (R=3) and m a margin, define the identity score as the negative hinge on post-attack cosine similarity,

$$\mathcal{L}_{\mathrm{id}}(\hat{I}) = -\frac{1}{R}\sum_{r=1}^{R}\max\Big(0,\ \cos\big(E_r(I),\ E_r(\hat{I})\big) - m\Big), \tag{3}$$

which reaches its maximum of zero once every recognizer is at or below m. Maximizing this score stops pushing faces already below the margin. The student minimizes

$$\mathcal{L}_{\mathrm{S}} = \big\|\delta_{\mathrm{S}} - \delta_{\mathrm{T}}\big\|_1 - \lambda_{\mathrm{fr}}\,\mathcal{L}_{\mathrm{id}}\big(\mathcal{A}(I + \delta_{\mathrm{S}} \odot M)\big), \tag{4}$$

$\delta_{\mathrm{S}} = G_\theta(I, M)$: the first term imitates the cached teacher, while the negative identity score penalizes recognizer similarity *after* a simulated attack. The student is rewarded for protection that survives editing rather than for fooling a clean recognizer. $\mathcal{A}$ is drawn from a differentiable simulator of nine families spanning reconstruction, compression, noise, geometry and a CLIP-conditioned (Radford et al., 2021) approximation of instruction edits (Brooks et al., 2023) (Appendix A.4). Which families it contains determines how robustness reads, so Section 5.6 splits results by modeled and unmodeled family.

3.4 Refinement, Training and Cost

At inference $\delta_0 = G_\theta(I, M)$ is sharpened by K projected-gradient steps (Madry et al., 2018),

$$\delta_{t+1} = \Pi_\epsilon\Big(\delta_t + \alpha\,\mathrm{sign}\big(\nabla_{\delta_t}\mathcal{L}_{\mathrm{id}}(\mathcal{A}(I + \delta_t \odot M))\big)\Big), \tag{5}$$

This is gradient ascent on the identity score, with $\alpha = 2.5\epsilon/K$. We use K=30 and start from δ_0 rather than from noise: over a zero start at the same K the learned initialization is worth $+0.0457$ $[+0.0368, +0.0550]$ on single portraits and $+0.0250$ on group photos (Section 5.6).

The student is trained on *Images of Groups* (Gallagher & Chen, 2009) (5,080 images, 28,231 faces) at 512×512 in three phases of $30/10/10$ epochs, each initialized from the previous phase's best checkpoint: teacher imitation (L_1 only), then the identity term through the simulator (λ_{fr}=0.2), then a higher identity weight at a low learning rate (λ_{fr}=0.5). Inference is not a single forward pass: the student rollout plus masked blending takes 9.8 s per image, and the deployed configuration, which adds refinement, takes 21.2 s on one RTX 5090.

4 EVALUATION PROTOCOL

Protocol choices in this literature are not neutral. Each control below moves the measured protection margin, and in every case the uncontrolled version reports a larger one, so a cross-method comparison is only as meaningful as the controls standing behind it.

4.1 THE PER-FACE CONTROLLED GAP

Whole-image evaluation of a group photo is confounded: a latent editor's encode–decode stage destroys the identity of small faces on its own, so a low post-edit similarity need not indicate a strong defense. We crop each detected face at margin 1.3, resize to 256×256, and run the identical attack along two paths, the protected image and an unprotected control of the *same* face. The headline quantity is the **protection gap** $\Delta = \mathrm{sim}_{\mathrm{unprot}} - \mathrm{sim}_{\mathrm{prot}}$, which attributes the measured effect to the protection rather than to the attack. We report every cell as *unprotected* $\rightarrow$ *protected (gap)*, because $+0.22$ means something entirely different when the control sits at 0.99 than at 0.40. The gap *ranks* methods whose objective is the identity; for anti-editing methods it is descriptive, and Section 5.4 scores those on the edit axis instead.

4.2 BOTH ATTACK ORDERINGS

Our protocol crops and resizes *before* attacking, and resampling destroys a perturbation spread over the whole frame, so the ordering is not neutral between method families. On the same 500 images under a VAE round-trip, scored by FaceLock's own optimization target, **crop-then-edit** gives gaps of $+0.585$ for Obscrowd, $+0.255$ for per-face PGD and $+\mathbf{0.043}$ for FaceLock against an unprotected control of 0.966, while **edit-then-crop** gives $+0.699$, $+0.280$ and $+\mathbf{0.574}$ against 0.985. The first moves FaceLock's gap $13\times$ while the face-localized arm barely moves; the second is what whole-image methods assume, and reporting crop-then-edit alone would credit us with a $14\times$ margin the protocol, not the method, produced. Every cross-method table therefore reports edit-then-crop as primary (Appendix B.2).

4.3 FOUR ADDITIONAL CHECKS

Four checks address ways an evaluation could overstate protection. Full procedures are in Appendix B.

(i) Seen versus unseen recognizers. FaceNet (Schroff et al., 2015) and two AdaFace (Kim et al., 2022) backbones guide Obscrowd and score Table 1. This favors us: the same models are unseen by the baselines. We therefore also test two evaluation-only CVLFace ViTs: `vit_kprpe` is FaceLock's target, while `vit_plain` is held out from both methods. We verify that both checkpoints distinguish genuine from impostor pairs; otherwise `transformers` 5.x can silently load untrained weights (Appendix A.5).

(ii) Recognizer-specific thresholds. A low similarity does not always mean verification fails. At FAR$=0.01$, 44,850 impostor pairs yield thresholds of 0.459, 0.221 and 0.161 for our three recognizers. A score of 0.35 can thus still verify, although a common 0.4 cutoff would count it as protected. Under reconstruction, the share judged non-matching falls from 40% to 22% on single portraits and from 73% to 33% on group photos. The threshold-free protection gap is unchanged; the practical verification claim is narrower.

(iii) Independent samples. Faces in one group photo share a scene and perturbation. We therefore bootstrap *photographs*, not faces (Efron, 1979; 10^4 resamples), and run paired tests at image level. Treating faces as independent makes intervals $1.82\times$ too narrow; the joint-versus-independent result is consequently stated as a bound.

(iv) Comparable perturbations. The realized L_∞ budget on $[0, 1]$ is 0.1600 for Obscrowd but 0.0151 for FaceLock; nominal ϵ values do not reveal this difference. We rerun FaceLock at our budget (Section 5.3) and compare at matched perceptual cost (Section 5.4). We also match perturbation density in configuration tests: an apparent $+0.087$ gain shrinks to $+0.012$ (p=0.081) after matching (Appendix C.11). Baselines use their authors' released code.

5 EXPERIMENTS

5.1 SETUP

Evaluation uses two held-out sets in full: 500 CelebA-HQ (Karras et al., 2018) single portraits (515 faces) and 326 WIDER FACE (Yang et al., 2016) group photographs (1,373 faces over 325; one yields no detection), plus ten native 512×512 portraits for the qualitative figure. Four analyses run on stated 250/150-image subsets at a fixed seed, the budget sweep (Section 5.4), the low-divergence split (Section 5.5), robustness and transfer (Section 5.6); threshold calibration and every cross-method comparison use the full sets. Four operations span reconstruction, compression and editing: a VAE round-trip, JPEG q=75, InstructPix2Pix (Brooks et al., 2023) with "add a smile", and SDEdit (Meng et al., 2022); InstructPix2Pix is a different model from the training surrogate, hence a held-out editor, and Section 5.6 adds 15 more. *Obscrowd* is the student at ϵ=0.16 with K=30. The four released baselines run from their authors' code at published settings: PhotoGuard, EditShield and FaceLock enter the editing evaluation (TM-1), while IDProtector is evaluated separately for identity-conditioned generation (TM-2; Section 5.3). *Per-face PGD* is our control: it shares our ϵ, mask and recognizers but has no learned initialization or simulator. For the post-edit identity metric, Obscrowd, per-face PGD and FaceLock are direct identity-targeting comparisons; PhotoGuard and EditShield target edit disruption and are included only to show its incidental effect on identity.

5.2 IDENTITY PROTECTION UNDER GENERATIVE EDITING

Table 1: **Post-edit identity under crop-then-edit.** Cells give cosine similarity to the clean identity (lower means less recognizable); parentheses show the gap from the matched unprotected edit (larger positive gap means more identity loss due to protection). Only the first method group targets identity; anti-editing methods are context, not ranked on their own objective. Bold marks the lowest identity-targeting similarity only when the unprotected edit retains identity. The shaded SDEdit controls have little remaining headroom, so those rows do not establish a method ranking. This table uses crop-then-edit and the recognizers used to optimize Obscrowd; Table 2 tests the baselines' ordering on a held-out recognizer.

		Identity-targeting			Anti-editing (context)	
Attack	**Unprot.**	**Obscrowd**	per-face PGD	FaceLock	PhotoGuard	EditShield
Single-portrait, 500 *images /* 515 *faces*						
VAE round-trip	0.974	**0.424** (+0.550)	0.741 (+0.233)	0.953 (+0.021)	0.912 (+0.062)	0.947 (+0.027)
JPEG q=75	0.997	**0.391** (+0.605)	0.713 (+0.284)	0.984 (+0.012)	0.955 (+0.042)	0.989 (+0.008)
InstructPix2Pix	0.711	**0.313** (+0.398)	0.549 (+0.162)	0.706 (+0.005)	0.687 (+0.024)	0.752 (−0.041)
SDEdit	0.310	0.172 (+0.138)	0.266 (+0.043)	0.291 (+0.018)	0.211 (+0.099)	0.260 (+0.050)
Multi-portrait, 325 *images /* 1,373 *faces*						
VAE round-trip	0.993	**0.328** (+0.665)	0.636 (+0.358)	0.959 (+0.035)	0.485 (+0.509)	0.934 (+0.060)
JPEG q=75	0.998	**0.320** (+0.678)	0.623 (+0.375)	0.964 (+0.034)	0.488 (+0.510)	0.937 (+0.061)
InstructPix2Pix	0.662	**0.287** (+0.375)	0.472 (+0.190)	0.654 (+0.008)	0.430 (+0.232)	0.633 (+0.029)
SDEdit	0.314	0.145 (+0.169)	0.249 (+0.065)	0.285 (+0.029)	0.140 (+0.174)	0.285 (+0.029)

Table 1 shows consistent identity reduction across the three attacks whose unprotected controls retain recognizable faces. Under InstructPix2Pix, the mean group-photo similarity falls from 0.662 unprotected to 0.287 with Obscrowd, a +0.375 protection gap; on single portraits it falls from 0.711 to 0.313 (+0.398). Under reconstruction and JPEG, group-photo gaps reach +0.665 and +0.678. Among the identity-targeting arms, Obscrowd has the largest gap in every informative row. Image-level paired comparisons favor it over per-face PGD and FaceLock on both sets (Table 9). PhotoGuard and EditShield appear only as context for residual identity, since this metric does not score their edit-blocking objectives.

The table also shows where to avoid overinterpreting a score. SDEdit alone reduces the unprotected control to 0.310–0.314, leaving little headroom to attribute identity loss to any protection. Operational verification depends on the recognizer: FaceNet is defeated on 59–91% of protected faces, while the two AdaFace models use stricter calibrated thresholds (Table 15 and Section 6). We therefore use the threshold-free protection gap for cross-method comparisons and state the recognizer for every thresholded claim.

Table 2: **Identity protection transfers to a held-out recognizer.** Single-portrait, 500 images, under the baselines' edit-then-crop ordering, scored by `vit_plain`, which is held out from both methods. Both arms target identity, so lower similarity is better for their shared objective.

Attack	Unprot.	Obscrowd	FaceLock
VAE round-trip	0.987	**0.292** (+0.696)	0.580 (+0.407)
JPEG q=75	0.994	**0.363** (+0.631)	0.976 (+0.019)
InstructPix2Pix	0.801	**0.247** (+0.554)	0.301 (+0.499)
SDEdit	0.565	0.249 (+0.316)	0.389 (+0.176)
mean gap		+0.549	+0.275

Mean gap, other arms: per-face PGD +0.218, PhotoGuard +0.141, EditShield +0.021. Paired differences against FaceLock, both recognizers and both sets, in Table 17.

5.3 Transfer and Matched Comparisons

The lead persists when both evaluation choices in Table 1 change. Under edit-then-crop on a recognizer held out from both methods, Table 2 gives a mean gap of +0.549 for Obscrowd versus +0.275 for FaceLock. Excluding the SDEdit row, whose unprotected control has limited headroom, gives +0.627 versus +0.308. On FaceLock's own optimization target, Obscrowd still leads by +0.125, although 107 of 500 images favor FaceLock (Table 17). FaceLock improves substantially under edit-then-crop, as expected for a whole-image method; the advantage we report is the one remaining after that ordering is restored.

Matched perturbation budget. On the full 500-image single-portrait set, Obscrowd leads FaceLock in the paired mean protection gap by +0.183 [+0.174, +0.192], with 484 images favoring Obscrowd (Table 21). We rerun FaceLock at our ϵ=0.16 bound, 10.7× its published face-region bound. Its whole-image run uses approximately 3.5× our realized mean perturbation, because Obscrowd is face-masked. The table reports all four operations, including InstructPix2Pix, where FaceLock has the lower protected similarity. This matched-budget comparison is scoped to single portraits.

The other threat model (TM-2). The multi-face formulation also matters for identity-conditioned generation, which uses a different attacker pipeline from editing. On the ArcFace (Deng et al., 2019) `glintr100` embedding used by InstantID (Wang et al., 2024) and IP-Adapter-FaceID (Ye et al., 2023), Obscrowd reaches 0.534 on group photos versus IDProtector's 0.776 (lower is better). IDProtector protects the largest conditioned face more strongly (0.523), but leaves the other faces at 0.827 because its preparation keeps one face. On single portraits, IDProtector is stronger: 0.148 [0.126, 0.170] versus our 0.379 [0.357, 0.402]. These results locate our gain in joint group protection rather than single-reference protection (Appendix C.3).

5.4 Protection and Perceptual Cost

The budget sweep shows that protection persists at substantially lower visual cost. On 250 single portraits after a VAE round-trip, Obscrowd achieves a +0.306 gap at face LPIPS 0.0302 (ϵ=0.04) and +0.437 at 0.0722 (ϵ=0.08; Table 22). Interpolation at per-face PGD's face LPIPS 0.046 suggests ≈+0.355 versus its full-set +0.233 (Figure 3 and Table 20). The sweep and PGD reference use different image counts, so this is a descriptive cost comparison, not a paired estimate.

At the deployed ϵ=0.16, protected InstructPix2Pix edits remain close to their unprotected counterparts: edit-LPIPS is 0.157 / 0.239 on single/group photos, below PhotoGuard's 0.187 / 0.399, while single-portrait edited-crop BRISQUE (Mittal et al., 2012) is 0.0 with Obscrowd versus 6.0 without protection. These perceptual proxies do not establish semantic edit success. The protection is visible: full-set face LPIPS is 0.144 / 0.216, compared with PhotoGuard's 0.156 / 0.412 and FaceLock's 0.033 / 0.044 (Table 20). On single-portrait VAE, PhotoGuard uses slightly more face-region energy (0.0425 versus 0.0403) but has a smaller identity gap (+0.062 versus +0.550), illustrating that magnitude alone does not explain this result.

Table 3: **Residual identity on low-divergence InstructPix2Pix edits.** Each arm retains its own lower half of edit-LPIPS divergence from the corresponding unprotected edit; this is a proxy for edit preservation, not a semantic success test or a paired subset across arms. Cells show post-edit identity similarity and, in parentheses, the share still verifiable by FaceNet at FAR=0.01 (τ=0.459). This analysis uses a fixed-seed subset of 250 single-portrait and 150 multi-portrait images (127 and 319 faces after each split). The appendix gives the scoring details in Appendix B.5. Lower values indicate less residual identity; anti-editing arms are contextual, not ranked on their edit-blocking objective.

	Single-portrait		Multi-portrait	
Arm	crop→edit	edit→crop	crop→edit	edit→crop
Identity-targeting methods				
Obscrowd	**0.332 (19%)**	**0.228 (6%)**	**0.286 (12%)**	**0.201 (4%)**
per-face PGD	0.625 (84%)	0.699 (88%)	0.517 (65%)	0.373 (34%)
FaceLock	0.794 (100%)	0.721 (94%)	0.730 (92%)	0.380 (34%)
Anti-editing methods (context)				
PhotoGuard	0.755 (98%)	0.843 (98%)	0.454 (51%)	0.198 (7%)
EditShield	0.791 (99%)	0.916 (99%)	0.703 (89%)	0.379 (35%)

5.5 Identity Protection When Edits Remain Close

Among single-portrait edits closest to their unprotected counterparts, Obscrowd leaves 19% of faces verifiable by FaceNet under crop-then-edit and 6% under edit-then-crop (Table 3). The corresponding FaceLock figures are 100% and 94%. PhotoGuard and EditShield, whose goal is to block edits rather than identity recognition, leave 98–99% verifiable; they provide context, not a head-to-head score of their own objective. On group photos under edit-then-crop, Obscrowd leaves 4% verifiable. The same ordering also lowers verification for whole-image methods, including PhotoGuard (51% to 7% across orderings), as full-resolution editing can degrade small faces even without effective protection.

The low-divergence subset is chosen separately for each method. Its values therefore measure conditional residual identity, not paired superiority or semantic edit success. The percentages in Table 3 are specific to FaceNet at FAR=0.01; the full-set per-recognizer and calibrated-majority outcomes are reported in Table 15.

5.6 Ablations and Robustness

Joint optimization preserves per-face strength at lower compute. Across 149 images stratified over all 15 WIDER scene categories, joint and independent optimization have nearly identical protection gaps: the paired difference is -0.0009 $[-0.005, +0.004]$ (p=0.413). Joint optimization uses 4.2× less compute on average and 9.7× less at ten faces. The result supports efficiency at comparable measured strength, not a claim that joint optimization is intrinsically stronger.

Learned initialization improves refinement. PGD from the student's output beats PGD from zero at the same step count by $+0.0457$ $[+0.0368, +0.0550]$ on single portraits (192 +/58 –) and by $+0.0250$ on group photos (101 +/49 –), and beats a random start by as much again; a random start is itself no better than zero, so the gain is learned structure rather than a non-zero start as such. Direct optimization is the *per-face PGD* arm of Table 1 (our ϵ, mask and recognizers, no student and no simulator), which we lead by $+0.242$ $[+0.232, +0.253]$ over 497 of 500 single portraits and $+0.225$ over all 325 group photos. Protection is flat in crowd size ($+0.430$ / $+0.401$ / $+0.416$ at N=2–3 / 4–7 / 8+), the property the joint formulation needs, and grows with face size (Appendix C.8).

Modeled transformations retain the strongest protection. Across the nine transformation families included in the simulator, the protection gap is $+0.664$ $[+0.647, +0.675]$; across three unseen families it is $+0.382$ (Table 18). Protection therefore transfers beyond the modeled set, but with a measurable reduction. Under affine re-alignment and DiffPure (Nie et al., 2022) at t=150, respectively, 92% and 95% of faces remain below the evaluated recognition threshold. These tests do not establish robustness to arbitrary adaptive transformations.

High-resolution portraits show the remaining opportunity. On ten native 512 × 512 portraits, InstructPix2Pix leaves 0.771 unprotected and 0.251 with Obscrowd. EditShield and FaceLock reach

still lower similarity (-0.003 and -0.028; Figure 4). This small qualitative set shows that our lead on the main portrait benchmark does not extend to every image scale.

Protection transfers to unseen editor architectures. Neither SDXL Img2Img (Podell et al., 2024) nor FLUX.1-schnell (Black Forest Labs, 2024) appears in training, the teacher or the simulator, all SD-1.5, and each has its own VAE and latent geometry. On 250 single and 150 multi images SDXL gives $0.604 \to 0.414$ ($+0.190$) and $0.562 \to 0.324$ ($+0.238$), FLUX $0.877 \to 0.540$ ($+0.337$) and $0.847 \to 0.417$ ($+0.430$): transfer across architecture, not across attack family (Appendix C.9).

6 Limitations

Dilution is the boundary of face-localized protection. The sharpest limitation is one mechanism with two consequences: **anything that reduces the protected face's share of the recognizer's input dilutes a face-localized perturbation.** Sweeping the attacker's crop margin on group photos under a VAE round-trip (325 images), our gap falls $+0.665/+0.521/+0.320/+0.108$ at margins 1.3 (deployed), 1.8, 2.5 and 3.5, against per-face PGD's $+0.358/+0.178/+0.073/+0.016$, while the unprotected control holds at 0.971–0.993 (Appendix C.7). At the widest crop our protected similarity has risen to 0.865 against that unprotected 0.973, and **the person is plainly identifiable again**. The attack-ordering result of Section 4.2 is the same mechanism from another angle: attacking before cropping reduces the multi-portrait gap from $+0.599$ to $+0.222$. This belongs to face-localized protection at small face sizes rather than to our method, retention being $100/78/48/16\%$ on group photos but a stable $100/90/89/89\%$ on single portraits. Re-identification from clothing, gait or context is out of scope for want of identity labels across images, and for the same reason we use *unlinkability* only for linkage by a face recognizer.

Protection under reconstruction and visual budget. Even when VAE reconstruction leaves unprotected identity largely intact, Obscrowd lowers mean similarity from 0.974 to 0.424 on single portraits and from 0.993 to 0.328 on group photos; JPEG shows similarly strong reductions (Table 1). VAE reconstruction attenuates high-frequency perturbations, making the budget consequential rather than eliminating the benefit. Across the tested budgets, a larger ϵ increases the protection gap while also increasing face LPIPS (Appendix C.6). We report this measurable trade-off between protection and visibility without claiming imperceptibility.

Protection is recognizer-dependent, and coverage is bounded. The gap is threshold-free; whether a protected face still *verifies* is not. At FAR$=0.01$ we defeat FaceNet on 59–91% of faces and the two AdaFace backbones, calibrated at 0.221 and 0.161, far less often, so any claim about what an attacker can no longer do has to name its recognizer and operating point.

7 Conclusion

Obscrowd jointly perturbs all detected faces in a photograph to reduce identity linkage after editing. On group photos edited with InstructPix2Pix under crop-then-edit, mean similarity falls from 0.662 to 0.287; joint protection matches independent per-face optimization at about $4.2\times$ less compute. The benefit depends on crop width and the recognizer, and adaptive attacks remain untested. Both the protection and its limits matter when deciding whether an edited photograph is safe to share.

AI Use Statement

We used large language models to **draft sections** of the main paper and appendix and to **aid and polish the writing**, including copy-editing and clarifying prose. This use was limited to writing. The authors reviewed and revised all AI-assisted text, checked the scientific claims and reported values against the underlying experiments, and take responsibility for the final content of the paper.

Ethics Statement

This is defensive technology, and we have tried to report it in a way that does not overstate what it delivers.

Dual use. The same machinery that suppresses biometric linkage of an edited photograph could be repurposed to evade recognition a subject has consented to. We therefore scope the objective narrowly, to the re-linkability of *edited outputs* by a face recognizer, and we do not evaluate, tune against, or claim any effect on authentication systems. The threat models of Section 3.1 state this scope explicitly.

An overstated defense is itself a harm. A user who believes a protection works when it does not is worse off than one who knows it does not, because they will publish an image they would otherwise have withheld. Every operational claim in this paper therefore names the recognizer and the operating point it holds for, rather than a single favorable threshold (Section 4.3 and Appendix B.3); we report the crop widths at which protection dilutes to nothing (Section 6), the recognizers against which identity is degraded rather than broken (Section 5.2), and the perceptual cost the protection charges (Section 5.4).

Data. All experiments use publicly available, consented imagery released for research: CelebA-HQ (Karras et al., 2018), WIDER FACE (Yang et al., 2016) and *Images of Groups* (Gallagher & Chen, 2009). No new imagery was collected, no human subjects were recruited, and no identity annotation beyond what those releases already carry was produced. We do not release protected copies of any individual's photograph. Re-identification from clothing, gait or scene context is out of scope and, as Section 6 states, our use of the word *unlinkability* is confined to linkage by a face recognizer.

Reproducibility Statement

The implementation, training curriculum and inference configuration needed to rebuild the method are given in Appendix A. The measurement conventions that a reproduction has to match, both attack orderings, per-recognizer threshold calibration, the cluster-bootstrap resampling unit and the residual-identity scoring conventions, are set out in Appendix B. Appendix E records which analyses run on the full held-out sets and which run on the stated fixed-seed subsets, together with the sample size behind every table, so that no result here is reported at an unstated n or pooled across sample sizes. The datasets are public and are named with their splits in Section 5.1 and Appendix A.6.

References

Namhyuk Ahn, KiYoon Yoo, Wonhyuk Ahn, Daesik Kim, and Seung-Hun Nam. Nearly zero-cost protection against mimicry by personalized diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Black Forest Labs. FLUX.1: Official inference repository for FLUX.1 models. https://github.com/black-forest-labs/flux, 2024.

Tim Brooks, Aleksander Holynski, and Alexei A. Efros. InstructPix2Pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

Ruoxi Chen, Haibo Jin, Yixin Liu, Jinyin Chen, Haohan Wang, and Lichao Sun. EditShield: Protecting unauthorized image editing by instruction-guided diffusion models. In *European Conference on Computer Vision (ECCV)*, 2024.

Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. ArcFace: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

Bradley Efron. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1): 1–26, 1979.

Andrew C. Gallagher and Tsuhan Chen. Understanding images of groups of people. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

Yuan Gan, Jiaxu Miao, Yunze Wang, and Yi Yang. Silence is golden: Leveraging adversarial examples to nullify audio control in LDM-based talking-head generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Xiao Guo, Xiufeng Song, Yue Zhang, Xiaohong Liu, and Xiaoming Liu. Rethinking vision-language model in face forensics: Multi-modal interpretable forged face detector. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Yuning Han, Bingyin Zhao, Rui Chu, Feng Luo, Biplab Sikdar, and Yingjie Lao. UIBDiffusion: Universal imperceptible backdoor attack for diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Zhenglin Huang, Jinwei Hu, Xiangtai Li, Yiwei He, Xingyu Zhao, Bei Peng, Baoyuan Wu, Xiaowei Huang, and Guangliang Cheng. SIDA: Social media image deepfake detection, localization and explanation with large multimodal model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of GANs for improved quality, stability, and variation. In *International Conference on Learning Representations (ICLR)*, 2018.

Hossein Kashiani, Niloufar Alipour Talemi, and Fatemeh Afghah. FreqDebias: Towards generalizable deepfake detection via consistency-driven frequency debiasing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Minchul Kim, Anil K. Jain, and Xiaoming Liu. AdaFace: Quality adaptive margin for face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Thanh Van Le, Hao Phung, Thuan Hoang Nguyen, Quan Dang, Ngoc N. Tran, and Anh Tran. Anti-DreamBooth: Protecting users from personalized text-to-image synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

Chanhui Lee, Yeonghwan Song, and Jeany Son. Data-free universal adversarial perturbation with pseudo-semantic prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Chun Tong Lei, Hon Ming Yam, Zhongliang Guo, Yifei Qian, and Chun Pong Lau. Instant adversarial purification with adversarial consistency distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Ziang Li, Hongguang Zhang, Juan Wang, Meihui Chen, Hongxin Hu, Wenzhe Yi, Xiaoyang Xu, Mengda Yang, and Chenjun Ma. From head to tail: Efficient black-box model inversion attack via long-tailed learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations (ICLR)*, 2018.

Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. SDEdit: Guided image synthesis and editing with stochastic differential equations. In *International Conference on Learning Representations (ICLR)*, 2022.

Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik. No-reference image quality assessment in the spatial domain. *IEEE Trans. Image Process.*, 21(12):4695–4708, 2012.

Andreas Müller, Denis Lukovnikov, Jonas Thietke, Asja Fischer, and Erwin Quiring. Black-box forgery attacks on semantic watermarks for diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Weili Nie, Brandon Guo, Yujia Huang, Chaowei Xiao, Arash Vahdat, and Anima Anandkumar. Diffusion models for adversarial purification. In *International Conference on Machine Learning (ICML)*, 2022.

Gaozheng Pei, Shaojie Lyu, Gong Chen, Ke Ma, Qianqian Xu, Yingfei Sun, and Qingming Huang. Divide and conquer: Heterogeneous noise integration for diffusion-based adversarial purification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: Improving latent diffusion models for high-resolution image synthesis. In *International Conference on Learning Representations (ICLR)*, 2024.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, 2021.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Ali Salar, Qing Liu, Yingli Tian, and Guoying Zhao. Enhancing facial privacy protection via weakening diffusion purification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Hadi Salman, Alaa Khaddaj, Guillaume Leclerc, Andrew Ilyas, and Aleksander Madry. Raising the cost of malicious AI-powered image editing. In *International Conference on Machine Learning (ICML)*, 2023.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. FaceNet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations (ICLR)*, 2021.

Yiren Song, Pei Yang, Hai Ci, and Mike Zheng Shou. IDProtector: An adversarial noise encoder to protect against id-preserving image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Hanhui Wang, Yihua Zhang, Ruizheng Bai, Yue Zhao, Sijia Liu, and Zhengzhong Tu. Edit away and my face will not stay: Personal biometric defense against malicious generative editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025a.

Junying Wang, Hongyuan Zhang, and Yuan Yuan. Adv-CPG: A customized portrait generation framework with facial adversarial attacks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025b.

Qixun Wang, Xu Bai, Haofan Wang, Zekui Qin, and Anthony Chen. InstantID: Zero-shot identity-preserving generation in seconds. *arXiv preprint arXiv:2401.07519*, 2024.

Zilan Wang, Junfeng Guo, Jiacheng Zhu, Yiming Li, Heng Huang, Muhao Chen, and Zhengzhong Tu. SleeperMark: Towards robust watermark against fine-tuning text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025c.

Song Xia, Yi Yu, Wenhan Yang, Meiwen Ding, Zhuo Chen, Ling-Yu Duan, Alex C. Kot, and Xudong Jiang. Theoretical insights in model inversion robustness and conditional entropy maximization for collaborative inference systems. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Long Xu, Jiakai Wang, Haojie Hao, Haotong Qin, Jiejie Zhao, and Xianglong Liu. Harnessing global-local collaborative adversarial perturbation for anti-customization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Shuo Yang, Ping Luo, Chen Change Loy, and Xiaoou Tang. WIDER FACE: A face detection benchmark. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. IP-Adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023.

Shuaiwei Yuan, Junyu Dong, and Yuezun Li. Where the devil hides: Deepfake detectors can no longer be trusted. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503, 2016.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

Fengfan Zhou, Bangjie Yin, Hefei Ling, Qianyu Zhou, and Wenxuan Wang. Improving the transferability of adversarial attacks on face recognition with diverse parameters augmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

APPENDIX

This appendix gives complete implementation and training details (Appendix A), full protocol details including both attack orderings, threshold calibration and the scoring conventions behind the low-divergence split (Appendix B), additional quantitative results (Appendix C), further qualitative material (Appendix D), and reproducibility details (Appendix E). Evaluation tables use the datasets and subsets identified in their captions; training, timing and implementation details are reported separately.

A IMPLEMENTATION AND TRAINING DETAILS

Deployment uses *only* the trained student and the refinement stage. The teacher is an offline component that produces distillation targets and is never invoked at inference.

A.1 STUDENT GENERATOR ARCHITECTURE

The student is a conditional U-Net in the Stable Diffusion latent space.

Table 4: Student generator. Base width C=64.

Component	Configuration
Input channels	16 = noisy latent (4) + image latent (4) + masks (8)
Encoder	3 stages, $C \to 2C \to 4C$, stride-2 downsampling
Bottleneck	residual blocks at $4C$ with timestep conditioning
Decoder	mirrored, skip connections from the encoder
Timestep embedding	sinusoidal, injected per residual block
Output head	pixel-space δ, bounded by $\epsilon \tanh(\cdot)$
Rollout	DDIM-style: 2 steps (training), 20 steps (inference)
Trainable parameters	$\mathbf{8{,}257{,}060}$

The two-step training rollout bounds the gradient tape, which must hold the SD VAE, the simulated attack and the full FR ensemble simultaneously. Inference is not memory-bound and uses the full 20-step rollout.

A.2 FACE LOCALIZATION, MASK ORDERING AND TILING

Faces are detected with MTCNN (Zhang et al., 2016). Each bounding box becomes a soft mask by Gaussian blurring, so the perturbation fades out at the mask edge and softens transitions at mask boundaries. On our multi-portrait set 46.6% of mask pixels lie strictly between 0 and 1, which is why our measured "off-face energy" (Table 20) is feathering partly attributable to mask feathering; the off-face percentages depend on the mask definition and should be interpreted with that convention in mind.

The student takes a fixed 8 mask channels. Faces are sorted by bounding-box area in *descending* order, and images with more than 8 faces are processed in tiles of 8, with tile outputs merged by a per-pixel maximum over the masked perturbations. The eight-channel input does not impose an eight-face cutoff; runtime and memory still grow with the number of tiles.

Why tiling matters. We ablate tiling by retaining only the first 8 masks while scoring all detected faces. In Table 5, later faces (zero-based index ≥ 8) have almost no protection: $+0.034$ versus $+0.416$ for earlier faces. Only 20 of 1,373 faces are affected. Deployed tiling processes them too, raising the multi-portrait gap from $+0.4052$ to $+0.4107$ and reducing contamination from -0.0189 to -0.0001. The aggregate gain is small because later faces are rare.

A.3 TEACHER

The teacher is a frozen Stable Diffusion 1.5 (Rombach et al., 2022). Per image it encodes to the latent space, runs a short DDIM (Song et al., 2021) trajectory, and at each step decodes the predicted clean latent to evaluate the guidance objective of Equation (2). The decoded difference from the original is

Table 5: **No-tiling ablation.** Only the first eight masks are processed here. Later-face results do not describe deployed Obscrowd, which tiles all detected faces. Face indices are zero-based; higher gaps mean more protection.

Face index	Mean gap	n
indices 0 to 7 (processed)	$+0.416$	80
index 8 or greater (omitted)	$+0.034$	20

clipped to the L_∞ ϵ-ball, multiplied by the union mask, and cached to disk as a $(3, 512, 512)$ tensor per training image. The teacher has no trainable parameters and is not part of the deployed system.

A.4 Differentiable Attack Simulator

Table 6: Differentiable attack families $\mathcal{A}$ available to the student loss and to the refinement stage. Every family is tested for gradient flow, non-triviality, and shape and range preservation before use.

Family	What it models
`vae_roundtrip`	encode–decode proxy for latent editing
SDEdit	noise-and-denoise editing (Meng et al., 2022)
`text_sdedit`	CLIP-conditioned (Radford et al., 2021) edit proxy
`jpeg_diff`	differentiable JPEG compression
`denoise_diff`	differentiable Gaussian denoiser
`geo_resize`	downscale-then-upscale
`geo_crop_jitter`	re-cropping with offset
`geo_rotate`	small in-plane rotation
`noise`	additive noise perturbation

Table 6 lists the families. Two of them, the differentiable JPEG and the differentiable denoiser, matter more than their share of the schedule suggests: substituting an additive noise proxy for compression, which is the cheaper option, makes JPEG the weakest attack in the whole evaluation, since the protection is then never optimized against the operation it is tested on. Both are active here, together with the three geometric families.

This composition is exactly why Section 5.6 splits robustness results by modeled and unmodeled family. A defense trained against a family and then evaluated on it is measuring optimization, not generalization, so pooled robustness results alone would obscure whether an operation was modeled.

A.5 Face-Recognition Ensemble

Table 7: Recognizers. The top block supervises training *and* produces the headline evaluation of Section 5.2, so those results are white-box for us. The middle block is used at evaluation only. The bottom row is the TM-2 metric.

Key	Backbone	Role
`facenet`	InceptionResnetV1 (Schroff et al., 2015)	train + eval
`arcface`	AdaFace iResNet-50 (Kim et al., 2022)	train + eval
`cvlface`	AdaFace iResNet-101 (Kim et al., 2022)	train + eval
`vit_kprpe`	CVLFace ViT, KPRPE	eval only; *FaceLock's target*
`vit_plain`	CVLFace ViT	eval only; held out from both compared methods
`glintr100`	ArcFace (Deng et al., 2019), antelopev2	TM-2 embedding metric

Two points about Table 7 bear directly on how fair the comparison is. First, our `cvlface` entry is an AdaFace iResNet-101 and is *not* FaceLock's optimization target; their target is the CVLFace ViT with KPRPE, so we add it as an evaluation-only recognizer rather than assume the comparison is already fair to them. Second, the identity objective and the headline evaluation share all three models in the top block, so those results are white-box for us; Section 4.3 is where that advantage is removed.

An untrained-weights hazard. In our evaluation environment, a CVLFace checkpoint-loading failure produced scores from untrained weights without raising an error. A self-identity check does

not detect this, because an untrained network also returns 1.0 for a pair of identical inputs. Every recognizer is therefore gated on a genuine-versus-impostor *separation* test before its scores are used: $+0.9926$ for `vit_kprpe` and $+0.9650$ for `vit_plain`. Loading multiple CVLFace checkpoints in one process additionally requires purging their shared module names from the interpreter between loads, or the second checkpoint silently reuses the first one's code.

A.6 Training Curriculum and Optimization

Table 8: Three-phase distillation curriculum (ϵ=0.16). Each phase is initialized from the best checkpoint of the previous phase.

Phase	Objective	λ_{fr}	Epochs	LR
1. Distillation	imitate teacher δ (L_1)	0.0	30	2×10^{-4}
2. FR fine-tune	add identity penalty via $\mathcal{A}$	0.2	10	5×10^{-5}
3. Aggressive FR	higher identity weight, low LR	0.5	10	1×10^{-5}

Training uses AdamW, mixed precision, gradient clipping at 1.0, and warmup-then-cosine learning-rate scheduling, on a single NVIDIA RTX 5090 (32 GB). The batch size is reduced whenever the CLIP-conditioned attacker is sampled, because its gradient tape (SD VAE, SD U-Net and the FR ensemble) dominates memory.

Training resolution. Training runs at 512×512. The 256×256 that appears throughout the evaluation sections is the *face-crop* size of the per-face protocol, and is a different quantity.

Training data. *Images of Groups* (Gallagher & Chen, 2009): 5,080 multi-portrait images, 28,231 faces. No evaluation image appears in training. The single-portrait evaluation set is therefore out of distribution for the student, which we return to in Appendix C.12.

A.7 Inference Configuration and Cost

The deployed configuration is the student at ϵ=0.16 followed by K=30 refinement steps with $\alpha = 2.5\epsilon/K$. Wall-clock, single RTX 5090:

Configuration	Seconds / image
student rollout + masked blend (all faces jointly)	9.8
deployed: student + K=30 refinement	21.2

Inference is not a single forward pass. It is a 20-step rollout plus, in the deployed setting, 30 refinement steps, and we report it as such.

B Protocol Details

B.1 Paired Comparisons for the Per-Face Protocol

Table 9: Image-level paired differences in protection gap behind Table 1: Obscrowd minus each comparator, with cluster-bootstrap 95% intervals and image-level sign counts. All numbers measure residual identity; PhotoGuard and EditShield do not optimize this objective, so their rows are descriptive rather than rankings of anti-editing success.

Comparator	$\Delta_{\text{ours}} - \Delta_{\text{arm}}$	**95% CI**	**Images +/–**
Single-portrait, 500 *images*			
per-face PGD	+0.242	[.232, .253]	497/3
FaceLock	+0.409	[.402, .416]	500/0
PhotoGuard (context)	+0.366	[.358, .374]	499/1
EditShield (context)	+0.412	[.405, .419]	500/0
Multi-portrait, 325 *images*			
per-face PGD	+0.225	[.218, .232]	325/0
FaceLock	+0.445	[.440, .451]	325/0
PhotoGuard (context)	+0.115	[.105, .126]	295/30
EditShield (context)	+0.427	[.421, .433]	325/0

B.2 Both Attack Orderings

Tables 10 and 11 give the full edit-then-crop results. Table 11 is worth reading carefully, because it is a case where a mean over attacks needs the unprotected column for interpretation. Across all four rows the mean gaps are +0.239 for Obscrowd and +0.228 for PhotoGuard, but three unprotected controls are already 0.115–0.395. JPEG retains identity in the unprotected control and gives gaps of +0.586 and +0.421, respectively.

The ordering effect, isolated. Table 12 isolates the mechanism on a single attack. A face-localized perturbation retains a substantial gap in both orderings, although its gap changes from +0.585 to +0.699. FaceLock changes more, from +0.043 to +0.574, consistent with an ordering-dependent effect of cropping and resampling.

B.3 Threshold Calibration

Thresholds are derived from 44,850 impostor pairs (Table 13). The two AdaFace models have much tighter impostor distributions than FaceNet, so their operating points sit far below the flat τ=0.4 that is common in this literature.

Table 14 shows that calibrated thresholds substantially reduce the reported non-match rate for two of the three recognizers. The threshold-free protection gaps are unchanged. The reported unprotected first-percentile similarities (0.941–0.982) indicate high control identity for the evaluated reconstruction setting; they do not establish complete separation between protected and unprotected score distributions. At FAR=0.01, the majority non-match rate is only 2% for VAE and JPEG on single portraits (Table 15), so operational verification claims need the recognizer and decision rule attached.

B.4 The Unit of Resampling

Confidence intervals resample *images*, carrying all faces of an image together, because faces in one photograph share a perturbation, a refinement trajectory and a scene. Resampling individual face-by-attack measurements instead treats 820 rows drawn from 40 images as 820 independent samples, and yields intervals 1.82× too narrow across all five summarizers here.

The choice of unit bites unevenly. An equivalence claim needs *more* power than a difference claim, and our joint-versus-independent result is an comparison, which is why Table 18 reports its observed difference and confidence interval. These data do not constitute a formal equivalence test without a prespecified margin. The difference claims in that table (learned initialization, direct PGD, ordering, robustness) hold at the image level, one of them unanimously across images.

Table 10: **Full cross-method comparison under the baselines' attack ordering** (edit-then-crop), single-portrait, 500 images, scored by the held-out `vit_plain`. This is the source table for Table 2. Shaded: the unprotected control has already lost most identity.

Attack	**Unprot.**	**Obscrowd**	PhotoGuard	per-face PGD
VAE round-trip	0.987	**0.292** (+0.696)	0.839 (+0.148)	0.714 (+0.274)
JPEG q=75	0.994	**0.363** (+0.631)	0.942 (+0.052)	0.701 (+0.293)
InstructPix2Pix	0.801	**0.247** (+0.554)	0.770 (+0.030)	0.586 (+0.215)
SDEdit	0.565	0.249 (+0.316)	**0.232** (+0.333)	0.475 (+0.090)
mean gap		+0.549	+0.141	+0.218

Attack	**Unprot.**	EditShield	FaceLock
VAE round-trip	0.987	0.963 (+0.024)	0.580 (+0.407)
JPEG q=75	0.994	0.985 (+0.010)	0.976 (+0.019)
InstructPix2Pix	0.801	0.845 (−0.045)	0.301 (+0.499)
SDEdit	0.565	0.472 (+0.093)	0.389 (+0.176)
mean gap		+0.021	+0.275

Table 11: **Limited identity headroom under multi-portrait edit-then-crop.** The unprotected similarity is already low for three operations after full-resolution processing of small faces. JPEG retains high unprotected identity and permits a clearer comparison of protection gaps.

Attack	**Unprot.**	**Obscrowd**	PhotoGuard	headroom
VAE round-trip	0.395	0.174 (+0.222)	0.121 (+0.274)	tiny
InstructPix2Pix	0.236	0.115 (+0.121)	0.097 (+0.139)	almost none
SDEdit	0.115	0.074 (+0.040)	0.036 (+0.079)	none
JPEG q=75	1.000	**0.413** (+0.586)	0.579 (+0.421)	**full**

Table 12: The same 500 single-portrait images, VAE round-trip, scored on `vit_kprpe`.

Ordering	Unprot.	Obscrowd	per-face PGD	FaceLock
crop-then-edit	0.966	+0.585	+0.255	**+0.043**
edit-then-crop	0.985	+0.699	+0.280	**+0.574**

B.5 Residual Identity on Low-Divergence Edits

Table 3 retains, for each arm and ordering separately, the faces whose protected edit has edit-LPIPS at or below that arm's median divergence from the corresponding unprotected edit. It reports the share of those faces a recognizer still verifies. The split is a perceptual proxy; it does not certify that the requested edit happened, and it need not select identical faces across arms. Two further choices determine the reported share.

Which recognizer. The three evaluation recognizers have different FAR=0.01 thresholds: 0.459, 0.221 and 0.161 (Table 13). Table 3 reports FaceNet at its own threshold. Table 16 additionally shows a descriptive mean of their raw scores compared with FaceNet's threshold. Because the score scales differ, that mean is not a calibrated multi-model verification rule. Per-recognizer and majority-vote outcomes on the full single-portrait set are reported separately in Table 15.

Which ordering. Crop-then-edit resamples the face crop before attacking, which discards much of a perturbation spread over the frame, so it measures whole-image methods where they are weakest (Appendix B.2). Both orderings are computed on the same faces in the same run, reusing the saved protected pixels; protection is not recomputed.

Table 16 reports the conditional subset under FaceNet and the descriptive score mean. On single portraits, Obscrowd's still-verifiable share is 19% / 6% under FaceNet and 6% / 2% under the score mean for crop-then-edit / edit-then-crop. The mean uses a threshold calibrated for FaceNet rather than for the averaged score, so only the FaceNet columns support a recognizer-calibrated verification statement. For calibrated outcomes across the full recognition ensemble, see Table 15.

The ordering separates the two settings. On single portraits the anti-editing arms sit at 98–99% under both orderings, so the conditional residual-identity result there does not depend on it. On group photos they recover sharply under their own ordering, PhotoGuard reaching 7% against our 4%,

Table 13: FAR-calibrated verification thresholds.

Recognizer	FAR=0.01	FAR=0.001
`facenet`	0.4590	0.6035
`arcface`	0.2209	0.3177
`cvlface`	0.1614	0.2332

Table 14: **What calibration changes.** Share of protected faces judged non-matching, under the flat τ=0.40 common in this literature and under per-recognizer FAR calibration. VAE round-trip on the full evaluation sets (500 single and 325 multi images; 515 and 1,373 faces). The aggregate falls because two of the three recognizers calibrate far below 0.40.

Set	**Recognizer**	flat τ=0.40	calibrated	Δ
Single	`facenet`	45%	63%	+18 pp
	`arcface`	33%	1%	−32 pp
	`cvlface`	43%	0%	−42 pp
	mean	**40%**	**22%**	**−19 pp**
Multi	`facenet`	71%	83%	+12 pp
	`arcface`	69%	9%	−60 pp
	`cvlface`	77%	6%	−71 pp
	mean	**73%**	**33%**	**−40 pp**

Table 15: **Verification outcomes at FAR-calibrated thresholds.** Percent of protected faces the recognizer no longer verifies, unprotected control in parentheses; single-portrait, full set (500 images, 515 faces). This is the table summarized in Section 5.2.

Attack	`facenet`	`arcface`	`cvlface`	**majority**
VAE round-trip	63 (0)	1 (0)	0 (0)	2 (0)
JPEG q=75	79 (0)	2 (0)	1 (0)	2 (0)
InstructPix2Pix	80 (9)	14 (2)	10 (2)	16 (2)
SDEdit	91 (65)	79 (29)	63 (17)	81 (30)

because a full-resolution attack degrades small faces without help from any defense. The residual-identity observation of Section 5.5 is scoped to single portraits for that reason.

C Additional Results

C.1 Cross-Model Paired Differences

We lead on all four cells, including both cells scored by the baseline's own optimization target. Table 17 also records that 107 of 500 single-portrait images favor FaceLock on its own target, which the aggregate hides and which we report.

C.2 Ablations in Full

The premise that faces in a group photograph barely overlap, which would make joint optimization trivially equivalent to independent optimization, is false on our data: soft masks overlap on 108 of the 149 images, on up to 56.0% of mask area.

Scaling with crowd size and face size is detailed in Appendix C.8.

C.3 Identity-Conditioned Generation in Full

IDProtector wins solo and loses multi, and the two right-hand columns say why: it drives the face it conditions on to 0.523 while the other faces in the same photograph stay at 0.827. Its preparation stage detects every face, sorts by area and keeps the largest, so the restriction is structural rather than a matter of configuration. PhotoGuard, a whole-image method, is the only baseline that degrades multi-portrait identity appreciably (0.593), and FaceLock leaves this axis essentially untouched (0.957), consistent with an objective aimed at breaking the edit rather than the identity.

Table 16: **Residual verification on low-divergence InstructPix2Pix edits.** Obscrowd has the lowest still-verifiable share in each of the eight displayed columns (lower is better). Each method is evaluated on faces at or below its own median edit-LPIPS divergence: 127 faces from 250 single-portrait images and 319 faces from 150 multi-portrait images after the respective splits. FaceNet uses its FAR=0.01 threshold of 0.459; *ens. mean* averages three recognizer scores and applies the same numerical threshold as an auxiliary score summary.

	Single-portrait				**Multi-portrait**			
	FaceNet		ens. mean		FaceNet		ens. mean	
Arm	c→e	e→c	c→e	e→c	c→e	e→c	c→e	e→c
Obscrowd	**19**	**6**	**6**	**2**	**12**	**4**	**6**	**1**
per-face PGD	84	88	80	88	65	34	69	18
PhotoGuard	98	98	98	98	51	7	44	2
EditShield	99	99	100	100	89	35	91	21
FaceLock	100	94	100	89	92	34	94	21

Table 17: Paired $\Delta_{\text{Obscrowd}} - \Delta_{\text{FaceLock}}$ under edit-then-crop, cluster bootstrap 95% CI over images with image-level sign test. `vit_kprpe` is FaceLock's own optimization target.

Set	**Recognizer**	**Difference**	**Sign test**
Single	`vit_kprpe` (theirs)	$+0.125$ $[+0.113, +0.137]$	393 +/107 −
Single	`vit_plain` (held out)	$+0.289$ $[+0.277, +0.300]$	495 +/5 −
Multi	`vit_kprpe` (theirs)	$+0.146$ $[+0.135, +0.158]$	314 +/11 −
Multi	`vit_plain` (held out)	$+0.143$ $[+0.133, +0.155]$	317 +/8 −

Table 18: **Ablations and robustness**, summarized in Section 5.6. Protection gap; intervals are cluster bootstrap 95% CIs over images, sign tests at image level. The first row gives a near-zero observed difference and its interval, not a formal equivalence result (Appendix B.4); the rest are difference results. Direct PGD is the same control as the *per-face PGD* arm of Table 1, so that comparison is given twice: the 25-image pilot, which carries the below-threshold contrast, and the full evaluation sets, whose test column reads as positive/negative images.

Comparison	**Effect**	**Test**
Joint − independent per-face	-0.0009	69 +/80 −
149 img / 637 faces	$[-0.005, +0.004]$	p=0.413
Learned init − zero init	$+0.0457$	192 +/58 −
250 single img, K=30	$[+0.0368, +0.0550]$	$p=6\times10^{-18}$
same, 150 *multi img*	$+0.0250$	101 +/49 −
Full pipeline − direct PGD	$+0.159$	25 +/0 −
below-τ 49–53% vs. 1%	$[+0.135, +0.190]$	$p=6\times10^{-8}$
same, on the full sets	$+.242$ / $+.225$	497/3; 325/0
robustness: modeled (n=9)	$+0.664$	$[+0.647, +0.675]$
unmodeled (n=3)	$+0.382$	92–95% below τ

C.4 Cost of Protection, in Full

Table 20 gives the full-set perceptual measurements. Obscrowd's face LPIPS is 0.144 / 0.216 on single/group photos, below PhotoGuard's 0.156 / 0.412 but above FaceLock's 0.033 / 0.044; we therefore make no imperceptibility claim. On single-portrait VAE, PhotoGuard has slightly greater face-region energy (0.0425 versus 0.0403) but a smaller identity gap (+0.062 versus +0.550). This example does not rank the anti-editing method on its own objective. Under InstructPix2Pix, Obscrowd's edit-LPIPS is 0.157 / 0.239, below PhotoGuard's 0.187 / 0.399; on single portraits edited-crop BRISQUE is 0.0 with protection versus 6.0 without it. These are perceptual proxies rather than semantic edit-success tests. The off-face column illustrates face localization: faces are roughly 9% of a WIDER frame, so a whole-image method spends about 91% of its budget where no recognizer looks. Our own off-face figure is soft-mask feathering, not misallocation, and is not comparable to theirs.

Table 19: **TM-2: the photo as an identity reference.** ArcFace (Deng et al., 2019) (`glintr100`) embedding similarity between the clean and the protected image, the axis InstantID (Wang et al., 2024) and IP-Adapter-FaceID (Ye et al., 2023) condition on; *lower is better* and there is no gap column because these are absolute. "Largest" is the face the pipeline conditions on. Underlined: a baseline beats us. Summarized in Section 5.3.

	solo	**multi**		
Arm	all↓	all↓	largest	others
IDProtector	**0.148**	0.776	0.523	0.827
Obscrowd	0.379	**0.534**	**0.359**	**0.569**
PhotoGuard	0.899	0.593	0.507	0.611
FaceLock	0.959	0.957	0.967	0.955

499 images / 584 faces (solo), 326 / 1,976 (multi): TM-2 uses the ArcFace detector InstantID conditions on, which finds more faces than TM-1's and none in one solo image. Solo CIs: $[0.126, 0.170]$ IDProtector, $[0.357, 0.402]$ ours.

Table 20: **What each method spends, and what it does to the attacker's edit.** Face LPIPS is the attacker's view of the perturbation; "off-face" is the share of perturbation energy outside the masks. BRISQUE (Mittal et al., 2012) is measured on the *edited output* (lower is more natural); edit LPIPS is the divergence between editing the clean image and editing the protected one. Lower edit LPIPS means greater perceptual agreement, not verified semantic edit success. Arms ordered by visibility; underlined marks where a baseline beats us. This is the source table for Section 5.4.

		what the method spends				*effect on the attacker's edit*		
Set	**Arm**	$\text{LPIPS}_{\text{face}}$ ↓	realized L_∞	$\overline{\lvert\delta\rvert}_{\text{face}}$	off-face %	BRQ_{IP2P} ↓	$\text{BRQ}_{\text{SDEdit}}$ ↓	$\text{LPIPS}_{\text{edit}}$
Single	*no defense*	–	–	–	–	6.0	23.6	–
	FaceLock	**0.033**	0.0157	0.0115	44.7	1.1	19.0	0.090
	per-face PGD	0.046	0.0383	0.0135	**0.2**	5.6	23.0	0.085
	EditShield	0.056	0.0353	0.0127	44.9	4.4	19.6	0.122
	Obscrowd	0.144	0.1600	0.0403	5.8	**0.0**	16.4	0.157
	PhotoGuard	0.156	0.1600	0.0425	10.4	0.9	**14.8**	0.187
Multi	*no defense*	–	–	–	–	56.6	53.4	–
	FaceLock	**0.044**	0.0157	0.0116	90.9	57.4	53.4	0.117
	EditShield	0.060	0.0353	0.0135	90.5	57.4	53.0	0.128
	per-face PGD	0.079	0.0598	0.0101	**15.8**	57.0	53.3	0.125
	Obscrowd	0.216	0.1599	0.0361	32.3	56.6	52.9	0.239
	PhotoGuard	0.412	0.1600	0.0477	42.2	**49.5**	**50.3**	0.399

Table 21: **Single-portrait matched-L_∞ comparison.** On the 500-image set, Obscrowd has the larger paired mean protection gap and leads on 484 images. FaceLock is rerun at our L_∞ over the whole image; its realized mean $|\delta|$ is 0.0829, versus 0.0122 at its published budget and approximately $3.5\times$ ours with our face mask. Results use the held-out `vit_plain` recognizer and edit-then-crop ordering. Shading marks the SDEdit control's lower unprotected identity.

Attack	**Unprot.**	**Obscrowd**	FaceLock @ our ϵ
Single-portrait, 500 *images*			
VAE round-trip	0.987	**0.292** (+0.696)	0.698 (+0.289)
JPEG q=75	0.994	**0.363** (+0.631)	0.854 (+0.140)
InstructPix2Pix	0.792	0.247 (+0.545)	**0.045** (+0.748)
SDEdit	0.567	**0.249** (+0.318)	0.288 (+0.279)
paired $+0.183$ $[+0.174, +0.192]$, 484 +/16 –, p=3.6×10^{-121}			

C.5 Matched Perturbation Budget on Single Portraits

Table 21 gives the single-portrait matched-budget result. Across the four operations, Obscrowd's paired mean protection-gap advantage is $+0.183$ $[+0.174, +0.192]$, with 484 of 500 images favoring Obscrowd. FaceLock has the lower protected similarity on InstructPix2Pix (0.045 versus 0.247); the reported lead is the paired aggregate over the stated set and conditions.

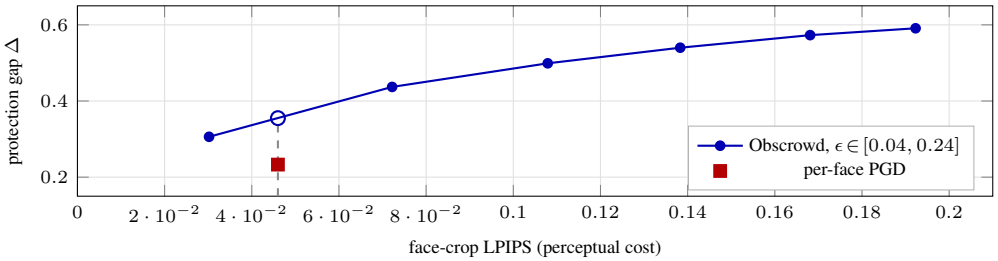

Figure 3: **Protection versus perceptual cost.** Obscrowd's VAE sweep uses 250 single portraits; the per-face PGD reference uses the full 500. Interpolating the sweep at PGD's face LPIPS 0.046 gives approximately +0.355 versus PGD's +0.233. This is a cross-sample, unpaired comparison. At ϵ=0.04, the measured gap is +0.306 at lower face LPIPS 0.0302.

Table 22: Budget sweep, single-portrait, VAE round-trip, 250 images. Unprotected similarity is 0.974 throughout, so the gap column is directly interpretable. It is plotted against perceptual cost in Figure 3.

ϵ	**Protected↓**	**Gap↑**	**face LPIPS↓**
0.04	0.668	+0.306	0.0302
0.08	0.537	+0.437	0.0722
0.12	0.474	+0.499	0.1079
0.16 *(deployed)*	**0.433**	**+0.540**	**0.1383**
0.20	0.401	+0.573	0.1681
0.24	0.383	+0.591	0.1923

C.6 Budget Sweep

The protection gap increases monotonically over the six tested ϵ values, with diminishing increments.

The marginal return falls monotonically. Gap per unit of face LPIPS across the sweep is 2.4, 1.1, 0.9, 0.7, 0.5, placing the knee at ϵ=0.08 and the deployed ϵ=0.16 past it. The ϵ=0.08 point offers a lower-visibility option on the tested single-portrait VAE sweep (+0.437 gap at face LPIPS 0.0722). The main tables use ϵ=0.16.

C.7 Crop-Margin Sweep, Both Settings

Table 23: Crop-margin sweep on group photos, VAE round-trip, 325 images, summarized in Section 6. Cells are protected similarity with the gap against the matched control; the shrinking gap is the dilution mechanism.

Margin	**Unprot.**	**Obscrowd**	per-face PGD	covers
1.3	0.993	**0.328** (+0.665)	0.635 (+0.358)	face
1.8	0.985	0.464 (+0.521)	0.807 (+0.178)	hair
2.5	0.971	0.651 (+0.320)	0.898 (+0.073)	shoulders
3.5	0.973	0.865 (+0.108)	0.957 (+0.016)	upper body
retention		100/78/48/16 %	100/50/20/5 %	

Table 24: The same sweep on single portraits, 500 images. The gap changes less than in group photographs over the tested margins.

Margin	**Unprot.**	**Obscrowd**	per-face PGD	FaceLock
1.3	0.974	0.424 (+0.550)	0.739 (+0.235)	0.953 (+0.021)
1.8	0.970	0.476 (+0.494)	0.766 (+0.204)	0.953 (+0.017)
2.5	0.969	0.479 (+0.491)	0.768 (+0.202)	0.953 (+0.016)
3.5	0.969	0.479 (+0.490)	0.768 (+0.201)	0.954 (+0.015)
retention		100/90/89/89 %	100/87/86/86 %	flat

Single portraits are stable under crop widening (100/90/89/89%) while group photos are not (100/78/48/16%). The reason is face coverage: a CelebA-HQ portrait's face already fills the frame,

so widening the crop adds background rather than un-protected identity signal, whereas a WIDER face occupies roughly 9% of the frame and there is a great deal of room to dilute.

C.8 Scaling with Crowd Size and Face Size

Table 25: Protection gap by crowd-size and face-size bins in the evaluated multi-portrait records. The crowd-size estimates are not monotonic.

Faces / image	Gap	Face size	Gap
N=2–3	+0.430	<32 px	+0.403
N=4–7	+0.401	32–128 px	*monotone*
$N\geq 8$	+0.416	≥128 px	+0.499

The three crowd-size bins show no monotonic decline (+0.430, +0.401, +0.416). The smallest and largest face-size bins differ (+0.403 versus +0.499), consistent with the face-coverage hypothesis, although the aggregated bins alone do not establish a causal mechanism.

C.9 Transfer to Unseen Editors, in Full

Table 26: **Transfer to editor architectures absent from training.** Neither model appears in the student's training, the teacher or the simulator, all of which use SD-1.5. This fixed-seed subset contains 250 single-portrait and 150 multi-portrait images (254 and 638 faces); both editors saw identical face sets. Summarized in Section 5.6.

Editor	Set	Unprot.↑	Obscrowd↓	Gap↑
SDXL Img2Img	single	0.604	0.414	+0.190
	multi	0.562	0.324	+0.238
FLUX.1-schnell	single	0.877	**0.540**	+0.337
	multi	0.847	**0.417**	**+0.430**

Each editor has its own VAE and latent geometry, and FLUX replaces the U-Net with a 12B transformer, so neither shares an architecture with anything the protection was optimized against. What they *do* share is the mechanism: noise-and-denoise editing is one of the nine families the simulator models, so this measures transfer across architecture rather than across attack family; it does not test an unmodeled operation.

FLUX leaves far more identity intact than SDXL (0.877 against 0.604 unprotected on single portraits), which is why the gap against it is larger: there is more to remove. This is the reason every cell here is reported as *unprotected → protected (gap)* rather than as a bare gap.

No baseline was run under these editors, so Table 26 carries no cross-method claim.

C.10 Learned Initialization and Refinement Tuning

Learned initialization improves refinement across the tested step counts. On the 250-image single-portrait subset, it beats a zero start at every K, from +0.030 at K=10 to +0.051 at K=50, and beats a random start by +0.044 to +0.050; all eight intervals exclude zero. On the 150-image multi-portrait subset at the deployed K=30, its gain over zero is +0.0250. A random start shows no consistent advantage over zero, with only a marginal advantage at K=50; the measured benefit comes from learned initialization.

The step-count sweep identifies a useful tuning interaction. Under the deployed $\alpha = 2.5\epsilon/K$ schedule, K=10 exceeds K=30 by +0.068 [+0.058, +0.079] on the 250-image single-portrait subset (201 +/49 –). Increasing attack draws per step from one to three leaves the comparison similar (+0.018 versus +0.015). Holding α fixed at 0.04 instead changes the reported step-count comparison from +0.0322 to −0.0105. Thus the observed K preference depends on the step-size schedule, and the two settings should be tuned together.

C.11 Protection and Perturbation Density

Across five tested (K, α) configurations, the protection gap closely tracks realized mean $|\delta|$ (Pearson $r{=}{+}0.935$). At matched runtime, three attack draws per step increase the gap by $+0.087$ while producing a 52% denser perturbation. After matching density at $\epsilon{=}0.105$, the difference is $+0.012\,[+0.000, +0.025]$ (sign-test $p{=}0.081$). Step-count comparisons also depend on α (Appendix C.10). We report realized density alongside L_∞ and runtime.

C.12 Is the Single-Portrait Setting Out of Distribution?

The student is trained only on group photographs, which raises the question of whether our weaker single-portrait numbers are an artifact of the student rather than a property of the task. Per-face PGD provides a useful comparison because it shares our ϵ, mask and recognizers but has no learned initialization or simulator.

Arm	**Single**	**Multi**	**Single − Multi**
Obscrowd (uses the student)	$+0.550$	$+0.665$	-0.115
per-face PGD (no student)	$+0.233$	$+0.358$	-0.125
PhotoGuard	$+0.062$	$+0.509$	-0.446
EditShield	$+0.027$	$+0.060$	-0.033
FaceLock	$+0.021$	$+0.035$	-0.013

Per-face PGD also has a lower gap on single portraits than on group photos (-0.125, versus -0.115 for Obscrowd). This weakens a student-only explanation for the difference, but it cannot exclude a student distribution effect: the datasets differ in face size, image statistics and face count, and these factors are not isolated here.

D ADDITIONAL QUALITATIVE RESULTS

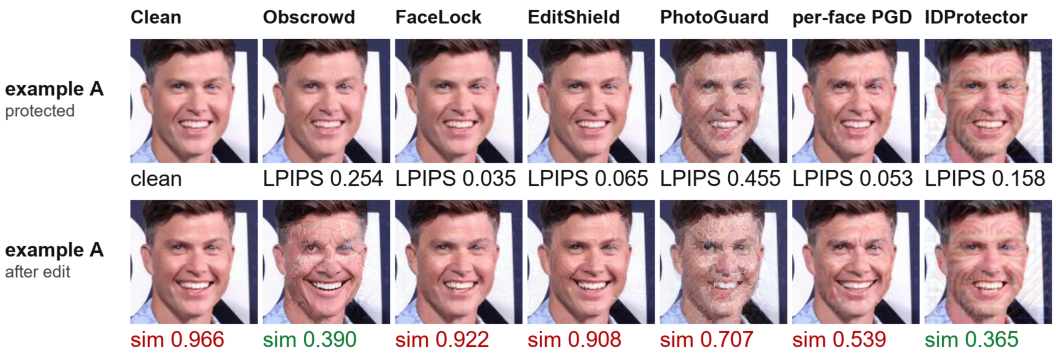

Figure 4: **Both halves of the trade, at native resolution**, summarized in Section 5.6. *Top:* what the user gives up (face-crop LPIPS). *Bottom:* what the attacker recovers after an InstructPix2Pix edit; green is below FaceNet's FAR=0.01 threshold of 0.459, red still verifies. Across the four selected examples, Obscrowd is $4/4$, IDProtector $1/4$, and the other arms $0/4$ under this FaceNet threshold. Drawn from ten native 512×512 portraits rather than from the group-photo set, whose median face is 32–64 px, so that the softness on display belongs to the perturbation and not to the source photograph. Per-image values are in Table 27; Figures 5 and 6 give all four examples across all six arms.

Table 27: **Per-image outcomes behind Figure 4.** Recovered FR similarity after an InstructPix2Pix edit. Bold marks a face below FaceNet's FAR=0.01 threshold of 0.459, i.e. verification defeated.

Img	Unprot.	**Ours**	FaceLock	EditShield	PhotoG.	IDProt.
1	0.732	**0.427**	0.824	0.851	0.769	0.542
2	0.877	**0.236**	0.926	0.925	0.652	0.649
3	0.700	**0.233**	0.722	0.874	0.634	0.542
10	0.966	**0.390**	0.922	0.908	0.707	**0.365**
defeated	0/4	**4/4**	0/4	0/4	0/4	1/4

Table 28: **Large-face set, full results** (10 native 512×512 portraits, 23 faces), same protocol as the tables of Section 5. IDProtector is included here as an editing-threat-model arm because this set is within the face size at which it can act.

Attack	Unpr.	**Ours**	PhotoG.	pf-PGD	EditSh.	FaceLock
VAE round-trip	0.983	**0.345**	0.745	0.684	0.950	0.958
JPEG q=75	0.998	**0.352**	0.820	0.652	0.978	0.987
InstructPix2Pix	0.771	**0.251**	0.665	0.560	0.775	0.800
SDEdit	0.345	0.168	**0.120**	0.270	0.314	0.307
IDProtector on this set: 0.684 / 0.676 / 0.555 / 0.257						
face LPIPS: FaceLock 0.041, pf-PGD 0.051, EditShield 0.068, IDProtector 0.166, **Obscrowd 0.169**, PhotoGuard 0.373						

Table 28 reports the large-face set in full. Two differences from CelebA-HQ are worth recording. Figure 7 illustrates seven prompt variations on one of the large-face portraits. **EditShield and FaceLock have slightly negative gaps under InstructPix2Pix** (-0.003 and -0.028) on this ten-image set; the protected means are marginally more recognizable than the unprotected mean. The small set does not support a general no-effect conclusion. And **IDProtector shows real editing-time protection here** ($+0.299$ under reconstruction, $+0.321$ under JPEG), comparable to per-face PGD, which it does not on the small faces of the multi-portrait set.

Why the qualitative figures are drawn from this set. Our multi-portrait benchmark has a median face size of 32–64 px. Cropping such a face and upscaling it for display produces a soft image no matter what the protection does, so a figure drawn from those images shows blur belonging to the *source photograph* rather than to the perturbation.

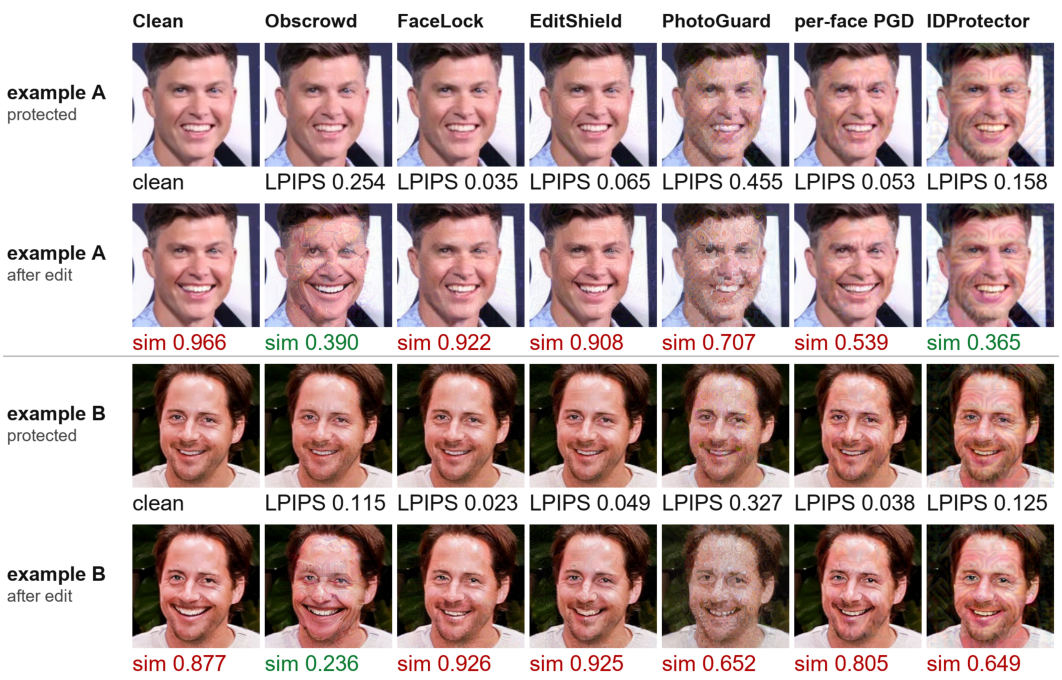

Figure 5: **Protection and appearance in two illustrated portraits.** Each protected crop is labeled with its face-crop LPIPS from the clean crop. In these examples, Obscrowd's values are 0.254 and 0.115, compared with PhotoGuard's 0.455 and 0.327; the other arms' values are labeled in the figure. After editing, Obscrowd's FaceNet similarities are 0.390 and 0.236, below the FAR$=0.01$ threshold of 0.459 in both cases. The protected crops show the visible texture accompanying this protection; full-set perceptual measurements are in Table 20. View at full resolution.

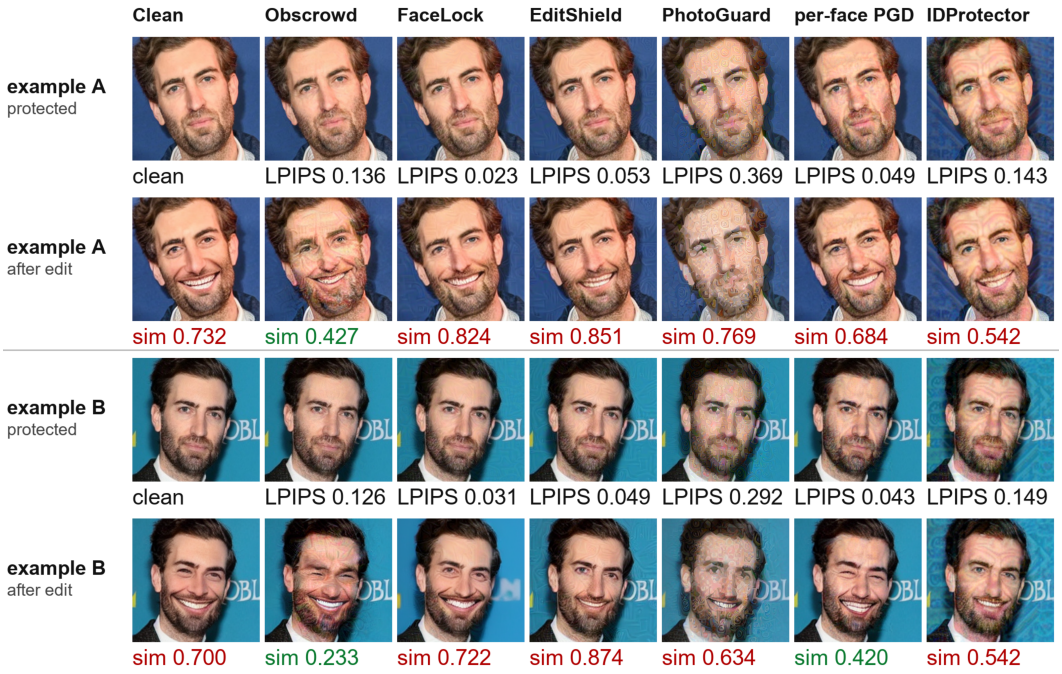

Figure 6: **Post-edit identity in two further examples.** The same faces after an InstructPix2Pix edit, labeled with the recovered FR similarity. Green marks a face below FaceNet's FAR=0.01 threshold of 0.459, that is, verification defeated; red marks a face that still verifies. Across the four examples in Figures 5 and 6, Obscrowd is the only arm that defeats verification on all four; per-face PGD and IDProtector manage one each; FaceLock none, its protected edits verifying at 0.72–0.93 although identity is its own stated target. EditShield and PhotoGuard also leave identity intact, but for them that is the contract rather than a failure of it, and PhotoGuard makes the distinction vivid: it visibly destroys the image in Figure 5 and the face still verifies at 0.63–0.77, which is the clearest single illustration of why an anti-editing objective is not a privacy objective. Per-image values are in Table 27.

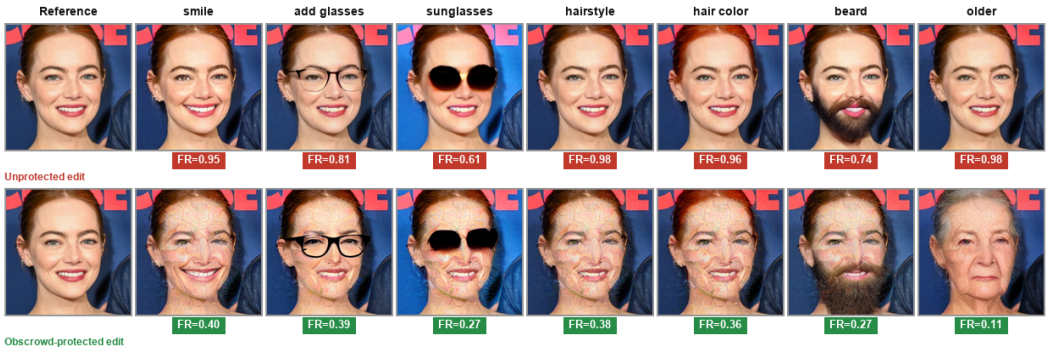

Figure 7: **Prompt variation on one large-face portrait.** For each of seven requested edits, the figure compares an unprotected output (red FR label) with an Obscrowd-protected output (green FR label) using the same reference face. The displayed similarities range from 0.61 to 0.98 without protection and from 0.11 to 0.40 with Obscrowd; the protected outputs also show visible texture. This figure shows one portrait across prompts; the large-face set results are in Table 28.

E REPRODUCIBILITY

The main TM-1 cross-method comparisons use the held-out set of 500 single-portrait images and 326 multi-portrait images, of which 325 produce detected faces for the per-face analysis. The experiment code and baseline versions should be checked against the eventual release before a reproducibility claim is made. Four analyses run on subsets of those same sets at a fixed seed, 250 single and 150 multi images: the budget sweep, the low-divergence split, robustness by operation family, and transfer to unseen editors. The ablations of Table 18 run at 149 images (joint versus independent on the 150-image multi-portrait subset, with one image yielding no detection), 250 single and 150 multi (learned versus zero initialization) and 25 (the direct-PGD pilot, superseded on the full sets by the *per-face PGD* arm of Table 1). The K curve of Appendix C.10 uses the 250-image single-portrait subset and a 40-image pilot on group photos, and is labelled as such where it is used. Statistical intervals are cluster bootstraps over images with 10^4 resamples. Wall-clock timings are measured on a single NVIDIA RTX 5090.